COVID-19: A Geospatial Analysis Using Brazilian Data

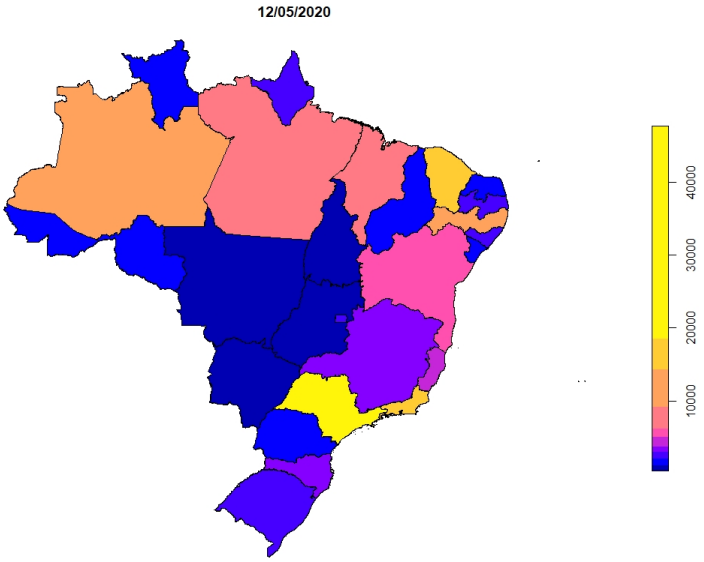
Hully Rolemberg

May 2020

I. Introduction

In late 2019, the World Health Organisation's (WHO) China office heard the first reports of an unknown virus causing a number of pneumonia cases in Wuhan, the capital of China's Hubei province. At that time, it was said that someone at the Huanan seafood market, in Wuhan, was infected with a virus from an animal. What started as an epidemic mainly limited to China, then became a global pandemic. By the beginning of May 2020, there were more than 4 million confirmed cases and 300 thousand deaths around the world. The so-called COVID-19 is an infectious disease that spreads primarily through droplets of saliva or discharge from the nose when infected people cough or sneeze. Thus, given the importance of social distancing to reduce the risks of infection and slow the pace of virus' spread, governments across the world implemented various stay-at-home orders. As a respiratory infectious disease, the virus spreads by clusters, but due to the lack of testing and sometimes failure to report on cases, there is a problem of underreporting in most of affected countries, in particular in Brazil.

FIGURE 1: NUMBER OF CONFIRMED CASES IN BRAZIL



With more than 10 million Brazilians living in densely populated favelas where usually there is no proper sanitation, and a great part of the labor force composed of informal workers, social isolation is very challenging in the country. Sao Paulo state is the epicenter of the COVID-19 in Brazil (see Figure 1) and it includes some of the cities with the highest population density. By the beginning of May 2020, there were more than 40 thousand reported cases in the state.

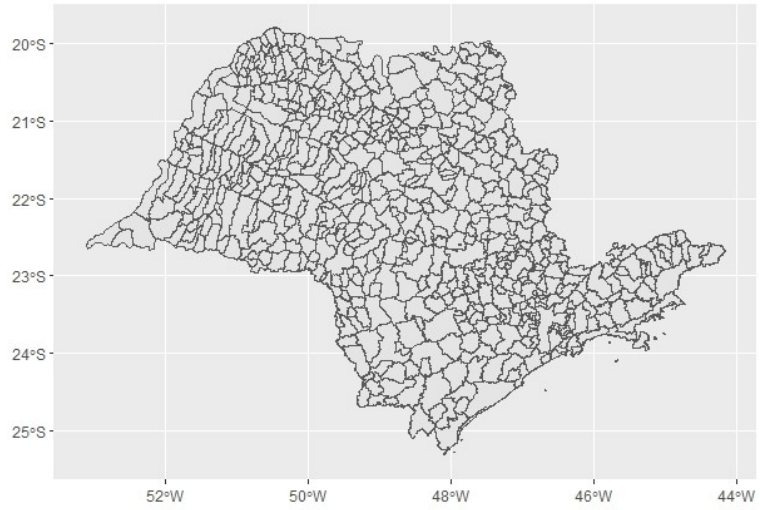
This paper evaluates the spreading pattern of the disease in Sao Paulo state using geospatial and social information so that I am able to predict the infection level in cities with no reported cases. To do so, I use the Kriging approach, which basically predicts the value of a function at a given point by computing a weighted average of the known values of the function in the neighbourhood of the point. I implement both usual Kriging and spatio-temporal Kriging. The results suggest that the geospatial analysis (although necessary) is not sufficient to explain the evolution of COVID-19 in the state of Sao Paulo.

II. Data

Data on COVID-19 is obtained from the State Health Secretaries (Portuguese: *Secretarias Estaduais de Saúde*). It includes updated municipal-level information on estimated population, number of reported cases and number of deaths since the first reported case, in February 25, 2020. Data is identified by the city name and the city identifier, provided by IBGE, the Brazilian Institute of Geography and Statistics. I also use cities' geometry and coordinates provided by IBGE. I merge the two datasetes using the IBGE identifier, so that all cities in Sao Paulo are observed at each day and those with no reported cases are filled with NAs. The sample period is from February 25 to May 12 (78 days).

There are 645 cities in Sao Paulo (see Figure 2), which represent around 45 million people. The most populated city is Sao Paulo and the most densely populated is Taboao da Serra.

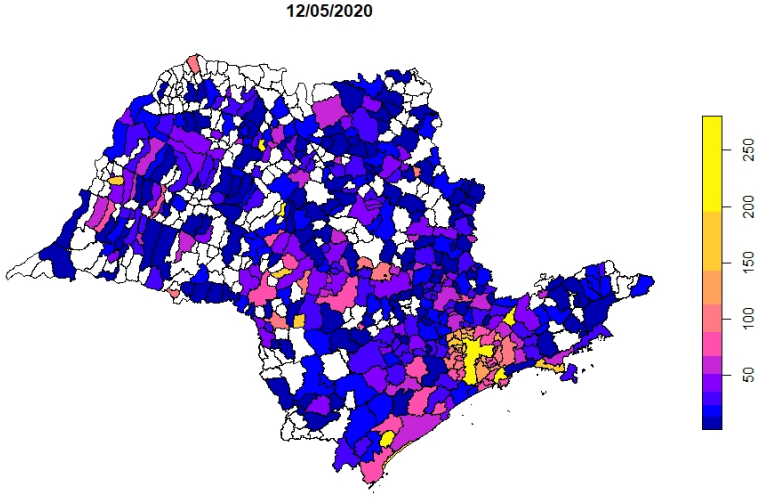
FIGURE 2: CITIES IN SAO PAULO



The first Brazilian case of COVID-19 was reported at the end of February in Sao Paulo city, which turned out to be the epicenter of the disease in the state. Then, the virus rapidly spread across the other cities. Figure 3 shows the number of cases per 100 thousand inhabitants, and we can see that there is indeed a cluster of cases around Sao Paulo city, but the disease has affected most of the state. Also, there are some cities with no reported cases, even though their neighbours do have infected people. In fact, due to the problems of testing and underreportation, there might be infected people even in the cities plotted in white.

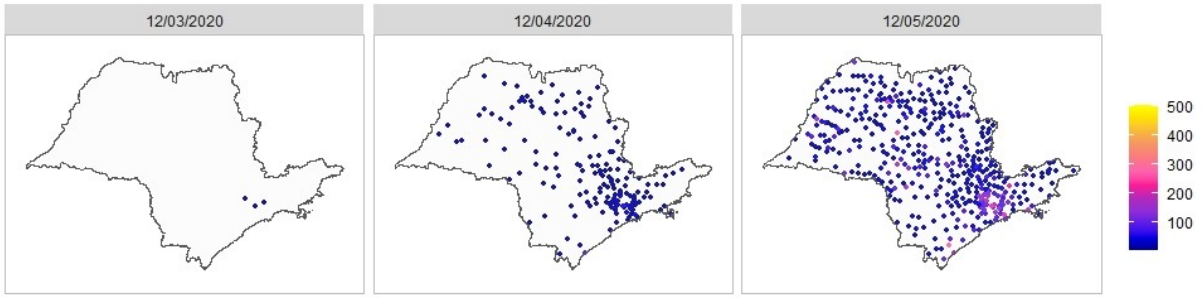
Figure 4 shows the spread pattern of the confirmed cases in the state during the sample period. As we can see, in the first half of March there were few reported cases in the state (44 in Sao Paulo city, 1 in Santana de Parnaiba, and 1 in Ferraz de Vasconcelos), but this scenario drastically changed

FIGURE 3: CASES OF COVID-19 PER 100K INHABITANTS IN SAO PAULO



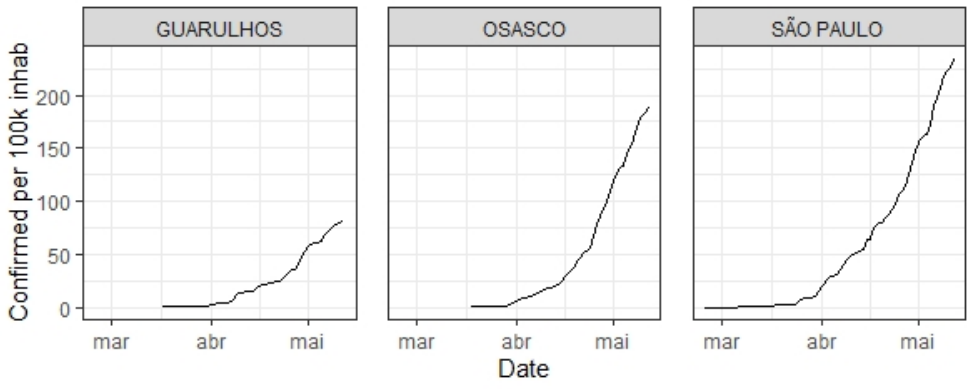
as the days went by. Again, we can observe that Sao Paulo city is the epicenter and the new cases seem to start from its surroundings. Two months after the first confirmed COVID-19 case, the virus has reached most of the cities in the state.

FIGURE 4: THE EVOLUTION OF CONFIRMED CASES OF COVID-19 IN SAO PAULO



Finally, Figure 5 shows the exponential shape of the epidemic curve in the three most affect cities in Sao Paulo (in terms of total confirmed cases). This is in accordance with what has been observed in other countries. The results do not show any stabilization pattern yet.

FIGURE 5: THE EVOLUTION OF CONFIRMED CASES PER 100K INHABITANTS IN THE THREE MOST AFFECTED CITIES IN SAO PAULO



III. Empirical Analysis

In this paper, my goal is to predict the presence of COVID-19 in cities that did not report any confirmed cases using a Kriging estimator. To do so, I first restrict my sample period to one single day, May 12 (which is also the most recent day in the sample). As we know, besides the underreportation problem, there is also a delay in the report of tests' results. So, if the Kriging forecasts are correct, we might see an approximation of the the Kriging forecasts and the observed numbers in the upcoming days.

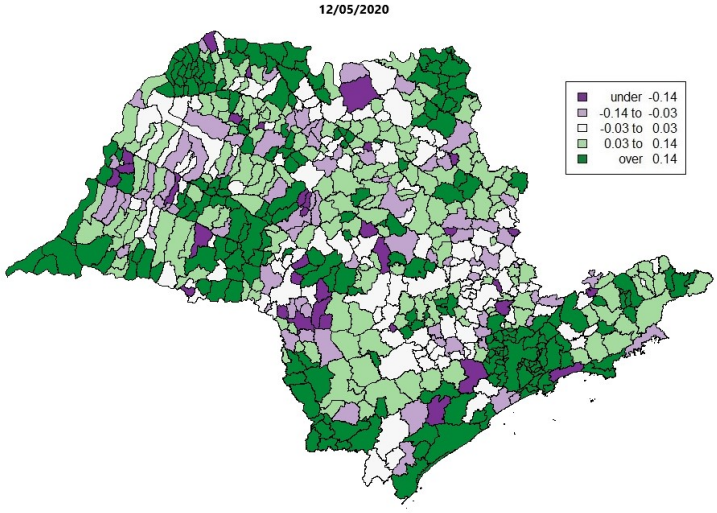
Since COVID-19 is an airborne disease, I expect it to spread from one city to its closest neighbours and so on. Thus, the level of spatial autocorrelation is a relevant aspect and it can be measured by Moran's I. However, the level of spatial dependency might not be the same across all places. So, to deal with that, we can decompose Moran's I accordingly to data contribution in each place and then identify local effects and potential "hotspots". The Moran's Local I is given by:

$$I = \frac{n}{\sum_i^n \sum_j^n w_{i,j}} \frac{\sum_i^n \sum_i^n w_{i,j}(z_i - \bar{z})(z_j - \bar{z})}{\sum_i^n (z_i - \bar{z})^2}$$

where z_i is some attribute of feature i, n is the total number of features and $w_{i,j}$ is the spatial weight.

In Figure 6, the Moran's Local I is used to measure spatial correlation of confirmed cases per 100 thousand inhabitants across the 645 cities in Sao Paulo. It is clear the existence of some correlation clusters in the state, specially in the surroundings of the capital. In fact, the higher level of spatial dependency between Sao Paulo city and its neighbours is justified by the existence of the "Greater Sao Paulo" (Portuguese: *Grande São Paulo*), a large metropolitan area where there is a densely populated urban core, sharing industry, infrastructure, and an integrated transport system.

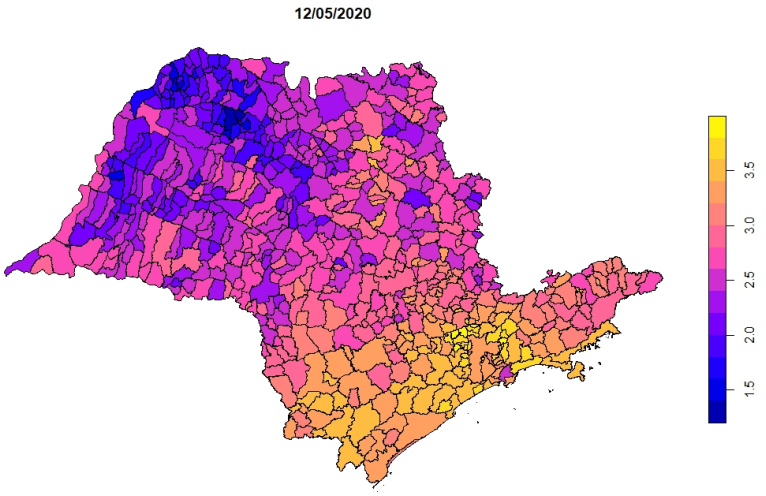
FIGURE 6: MORAN'S LOCAL I - CONFIRMED CASES PER 100K INHABITANTS



So far, I have focused on the number of confirmed cases per 100 thousand inhabitants, and I do the same in my econometric analysis. I first estimate an Ordinary Kriging and then two models of Universal Kriging. Model 1 includes the distance of each point to Albert Einstein Hospital (the first Brazilian hospital to report an infected person), as a covariate. Model 2 includes distance to hospital and the percentage of people that live in houses with density above 2% - where *density* is defined as the ratio of total residents to the number of rooms in the house, excluding bathrooms. Again, since COVID-19 is an airborne disease, more densely populated houses are more likely to be an infection

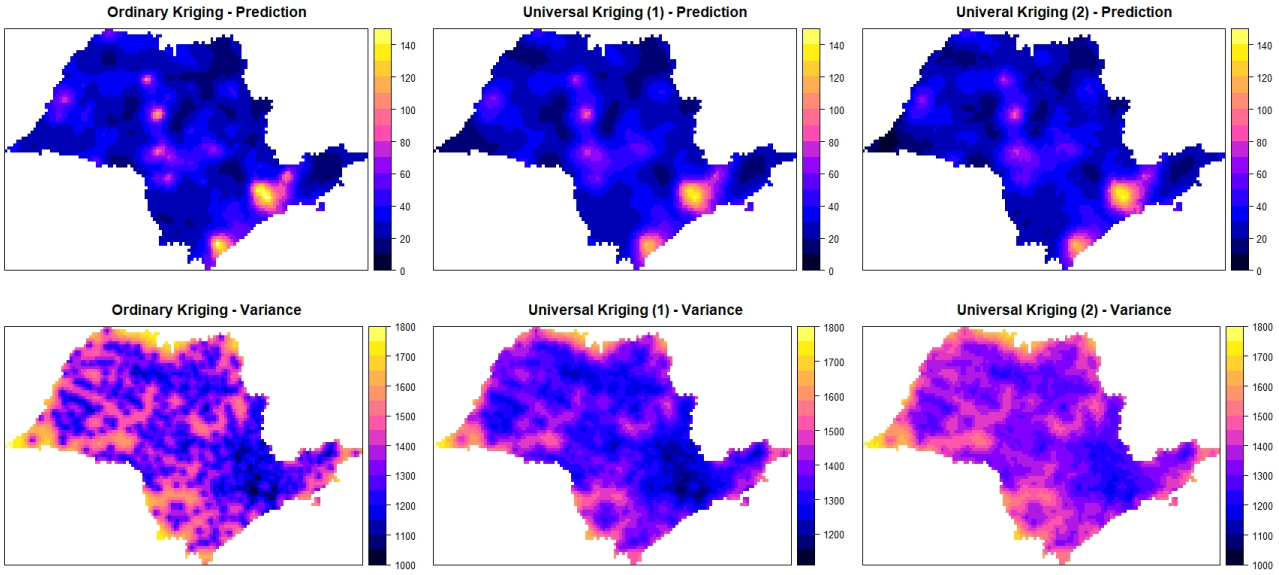
cluster, so I also use it as a covariate. Figure 7 shows the distribution of the density variable in the state.

FIGURE 7: PERCENTAGE OF PEOPLE THAT LIVE IN HOUSES WITH DENSITY ABOVE 2% IN SAO PAULO



Kriging predictions and their estimated variances are presented in Figure 8. We can see that there is no big difference between the prediction results generated by Ordinary Kriging and Universal Kriging. We can also observe the existence of some *hotspots* in the countryside, showing the disease progression in the interior of the state, where the healthcare systems are overall less prepared to respond to the virus. Unfortunately, the estimated variance is large in both Ordinary and Universal Kriging, implying that the predictions are not accurate. Implementing the cross validation (LOOCV), we deduce that Model 2 provides the best fit ($RMSE_{ordinary} = 37.096$, $RMSE_1 = 36.561$, $RMSE_1 = 36.03$).

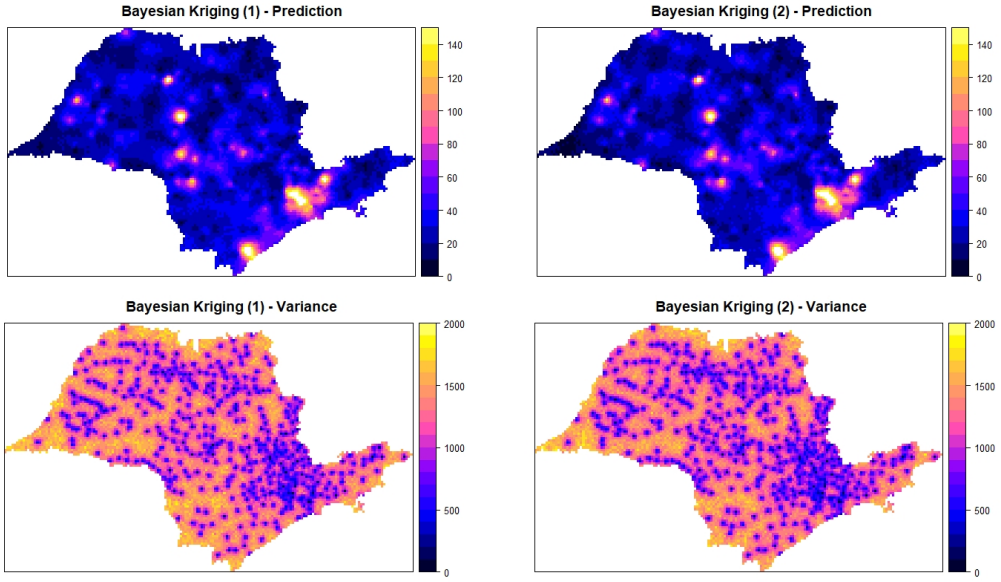
FIGURE 8: KRIGING INTERPOLATION ON MAY 12, 2020.



Alternatively to the usual Kriging estimation, I also implement the Bayesian Kriging interpolation. The results are shown in Figure 9. The Bayesian estimation produces more intense *hotspots* of the

disease in the state, which is reflected by the lighter colors in the map (the prediction scale is the same as in the previous estimation). However, the results are even less accurate than in the previous estimation. In sum, the results are very similar for both usual and Bayesian Kriging.

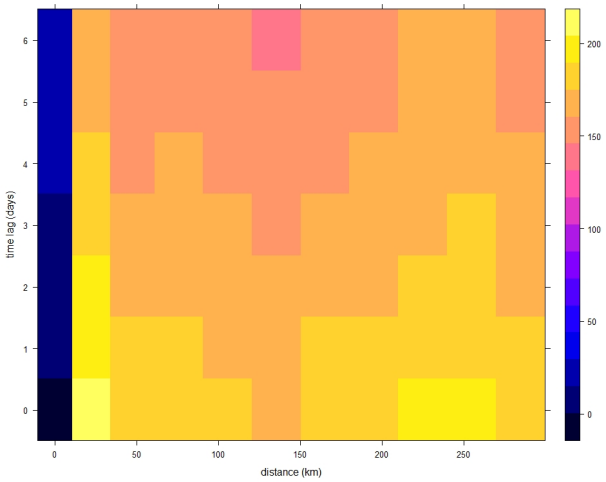
FIGURE 9: BAYESIAN KRIGING INTERPOLATION ON MAY 12,2020.



Up to this point, spatial correlation did not properly explain the spread of COVID-19 in Sao Paulo. This might be due to (i) the insufficient number of tested people, and (ii) the delay in the report of confirmed cases. Also, it should be included in the universal Kriging regression a measure of mobility across the cities, so it would be possible to map how the virus has been moving within the state.

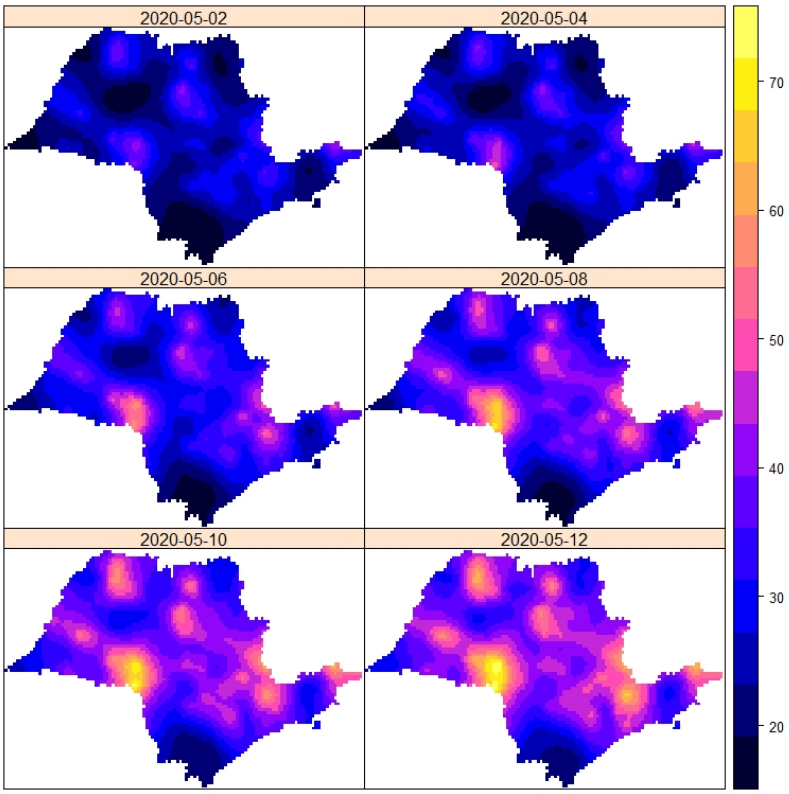
Observe that, the previous analysis was based on data collected on May 12, but information on infected people in Brazil is available since the first reported case, in February. So I can also implement a spatio-temporal analysis. Now, I restrict my sample to the period from February 25 to May 12. I also use the number of confirmed cases per 100 thousand inhabitants as dependent variable. First, I estimate the spatio-temporal semivariogram, but the results do not suggest any clear structure of spatio-temporal dependency in data (see Figure 10).

FIGURE 10: SPATIO-TEMPORAL EMPIRICAL SEMIVARIOGRAM



If I still proceed with my analysis and calculate the spatio-temporal Ordinary Kriging estimator, the results are actually inconclusive. Figure 11 shows the predictions for some of the most recent days in the sample period. We can observe a rapid progress of the disease in the state as days go by, but differently from what we observed in the previous estimations, the epicenter is no longer around Sao Paulo city. Instead, the main *hotspot* in the map is close to Assis and Marilia, two cities that reported less than 30 cases per 100 thousand inhabitants during the whole sample period. Besides that, the scale in Figure 11 is much lower than the one of static predictions. The estimated variance is also inconclusive (see the Appendix A.14). Spatio-temporal Kriging was implemented in *R* using the *krigeST* function from the *gstat* package.

FIGURE 11: ORDINARY SPATIO-TEMPORAL KRIGING: PREDICTIONS



Even though I could have used data up to May 26, there was a large acceleration in the number of new COVID-19 cases from May 12 to May 16, so that the model was not converging. Since I had to fix a sample to implement the empirical analysis, I chose the safest one.

A Appendix

FIGURE A.12: SEMI-VARIOGRAM (12 MAY, 2020)

(A) Ordinary Kriging (B) Model 1 (C) Model 2

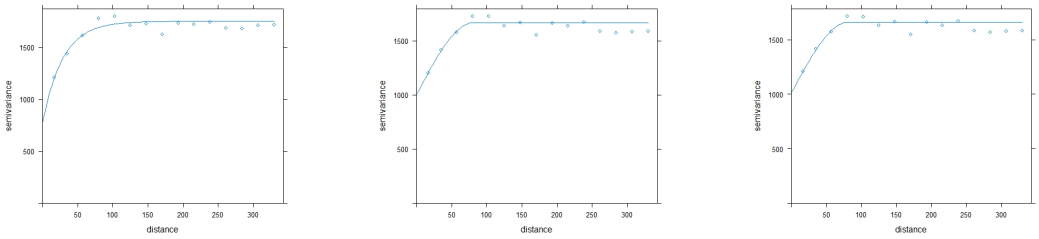


FIGURE A.13: BAYESIAN KRIGING - PRIORS

```
n.samples <- 1000
starting <- list("phi"=3/0.5, "sigma.sq"=50, "tau.sq"=1)
tuning <- list("phi"=0.1, "sigma.sq"=0.1, "tau.sq"=0.1)
tuning <- list("phi"=1, "sigma.sq"=0.1, "tau.sq"=0.1)
priors <- list("beta.Flat", "phi.Unif"=c(3/1, 3/0.1),
               "sigma.sq.IG"=c(2, 5), "tau.sq.IG"=c(2, 5))
cov.model <- "exponential"
```

FIGURE A.14: ORDINARY SPATIO-TEMPORAL KRIGING: PREDICTION VARIANCE